STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:3055–3071 (DOI: 10.1002/sim.1545)

Estimating regression models with unknown break-points

Vito M. R. Muggeo*,†

Istituto di Statistica Sociale, Scienze Demografiche e Biometriche, Facoltà di Economia, Università di Palermo, 90121 Palermo, Italy

SUMMARY

This paper deals with fitting piecewise terms in regression models where one or more break-points are true parameters of the model. For estimation, a simple linearization technique is called for, taking advantage of the linear formulation of the problem. As a result, the method is suitable for any regression model with linear predictor and so current software can be used; threshold modelling as function of explanatory variables is also allowed. Differences between the other procedures available are shown and relative merits discussed. Simulations and two examples are presented to illustrate the method.

KEY WORDS: break-point; threshold value; segmented regression; non-linear model; Taylor expansion

1. INTRODUCTION

In common regression analysis 'the response' (that is, the left-side of the regression equation) is modelled as a linear function of the explanatory variables $X_1, X_2 \ldots$. Thus, for instance, the log-expected value for Poisson regression, the logit transformation for binomial data, and the log hazard function for the Cox model are expressed as a linear combination of regressors. Sometimes it may happen that the relationship between the response and some explanatory variables is non-linear, showing a few values where the effect on the response changes abruptly. These values are called break-points, change-points, transition-points or switch-points [1, 2]; the word 'threshold' is also used [3], but some authors use it just when there is no effect on the response before of such a value [4, 5]. In toxicology the 'no-observed-effect-level' or 'no-effect-concentration' are strictly related to meaning threshold value [6, 7]. Non-linear relationships with break-points are said to be piecewise, segmented, broken-line or multi-phase regressions [8–11].

In what follows no distinction is made between the different words; anyway the effect on the response changes before and after the break-points where the regression function is continuous, but first derivatives discontinuous. Figure 1 illustrates some possible segmented regression between the response and generic covariate Z, say.

*Correspondence to: Vito M. R. Muggeo, Dipartimento di Metodi Quantitativi per le Scienze Umane, Università di Palermo, 90121 Palermo, Italy
†E-mail: vito.muggeo@giustizia.it

Received January 2002
Accepted March 2003

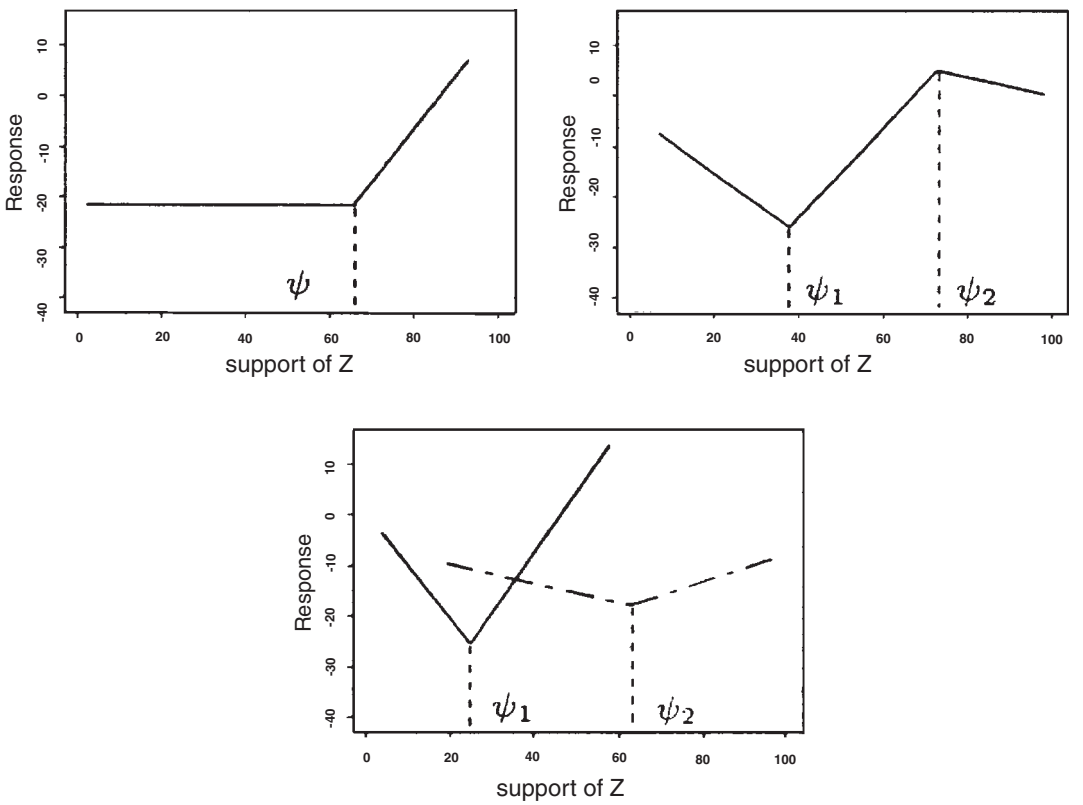

Figure 1. Some possible segmented relations: single relationship with one break-point; single relationship with two break-points; relationship depending on a dichotomous variable.

In analysing such 'kink-relationships', one is usually interested in the break-point location and in the relevant regression parameters (that is, slopes of straight lines). The classical methods used to take into account non-linear effects, such as polynomial regression, regression splines and non-parametric smoothing, are not suitable because the change-points are fixed *a priori* (regression splines) or are not considered at all (smoothing splines and polynomial regression). Furthermore, regression parameters obtained in regression splines or polynomial regression approach are not directly interpretable [12].

Thus, in order to estimate the model and to get meaningful parameters, alternative procedures have to be employed.

In biomedical applications it is easy to see the effect of some risk factor on the response to change before and after some threshold value. For instance, in mortality studies, the relationship death–temperature is V-shaped, so it may be of interest to estimate the optimal temperature where the mortality reaches its minimum [13]. In the Stanford heart transplant data, transplanted subjects aged around 50 or more years have a major risk with respect to younger subjects [3, 12], and the effect of dust concentration on bronchitis in German workers is significant for high values of exposure [14, 15]. Many other examples with piecewise terms

have been studied in the literature including mouse leukaemia [16], small-cell lung cancer [3] and time series of AIDS cases [1]. Moreover, threshold values may be useful to classify subjects in (generally two or three) groups.

When the break-point parameters have to be estimated, standard likelihood-based inference is complicated by the fact that the log-likelihood is just piecewise differentiable and so the classical regularity conditions are not met [2, 8, 17]: moreover, break-points are non-linear parameters and standard maximization procedures cannot be used [2]. Thus, to estimate such models, not only non-standard optimization techniques have to be handled, but some other additional techniques should be performed, see Section 3. These difficulties are actually limits for the usual employment of broken-line models in practice.

In this paper an alternative approach to estimate broken line models is suggested, reducing the problem to a linear framework. The method is shown to be conceptually very simple but also very general and suitable for several situations; multiple parameters are easily accounted for and for estimation only the starting values for the break-points are requested.

The organization of this paper is as follows: Section 2 introduces two well known data sets in the break-point estimation context that will be analysed in Section 6; Section 3 briefly reviews alternative methods usually found in the literature; in Section 4 the old original idea of Box and Tidwell to fit generic non-linear terms is illustrated and then adapted to segmented regression, discussing some simplifications that can occur; in Section 5 results from some simulations are presented, and finally Section 7 contains some discussions and conclusions.

2. EXAMPLES

In order to make useful comparisons, two routinely used data sets in segmented regression analysis have been used to apply the method proposed in this paper; analysis and results are presented in Section 6.

2.1. Survival analysis: Stanford heart transplant data

Various versions of this data set have been analysed by several authors for different aims. Roughly speaking, data consist of survival of patients on the waiting list for the Stanford heart transplant programme begun in October 1967, along with subject-specific covariates, including age, sex and several prognosis factors.

Following Molinari *et al.* [3], data used here refer to $n = 157$ patients who had received a heart transplant by February 1980; of these, 102 were deceased and 55 were alive; age at transplant and the prognosis variable 'mismatch score' (t5) are available. For useful reference, this data set is also supplied in the `survival5` library for the `R` software as 'stanford2' and concerns just subjects with non-missing values for the variable t5.

The aim is to assess whether the failure time depends on age at transplantation (`AGE`), and whether such a possible risk is constant along the entire range of such an explanatory variable.

2.2. Logistic regression: chronic bronchitis and dust concentration data

The problem here is to investigate whether concentration of cement dust in the workplace air is a possible risk factor of chronic bronchitis. This is a classical problem in occupational epidemiology, where the interest lies in assessing the threshold value of some chemical

substance, that is, the maximum concentration under which no influence on the employee's health is observed.

In Germany, a multi-centre study between 1972 and 1977 enrolled a cohort of workers from a dusty plant to study the relationships between health and cement concentration. By means of several measurements made directly at the workplace and questionnaires finalized to obtain medical data, $n = 1246$ observations were available including data on occurrence of chronic bronchitis (binary outcome variable, yes/no), dust concentration (DUST, continuous variable expressed in mg/m^3), smoking status (SMOKE, binary variable smoker/non-smoker) and duration of dust exposure (EXPO, continuous variable expressed in years). These data have been discussed and analysed elsewhere [4, 14, 17]; following these authors DUST will be analysed as $\log_{10}(\text{Dust} + 1)$.

Thus the main question is: adjusting for the other confounding variables, is DUST associated with chronic bronchitis? Does a threshold value exist?

3. METHODS DEALING WITH SEGMENTED REGRESSION

In the literature several methods have been proposed to handle regression models with unknown break-points. Those usually found can be substantially grouped into:

1. Methods that separately estimate the change-points. If the break-point is fixed, the models are usual linear ones, without any problems of estimation and inference; the break-point may be calculated by a simple inspection of the scatter smooth plot [13, 18] or by some specific algorithm, maybe exact or grid-search type [1, 4, 10, 16, 19, 20]. Since the estimation of beta parameters is conditioned on the break-point value obtained beforehand separately, this is not handled as a completely true parameter, estimation of the model occurs assuming the break-point to be known, and as a consequence the number of degrees of freedom is incorrect and the full covariance matrix of estimates is not available.
2. Methods that modify the response probability distribution in order to include the threshold parameter. For instance, to fit a threshold value in a logistic model with just one explanatory variable, Cox [5] introduces a particular probability distribution for the response, assuming no effect before the threshold value. These methods, as the same author discusses, are poor and difficult to generalize.
3. Methods that approximate the segmented relationship through a continuous differentiable function on the overall range of the explanatory variable [11, 21, 22] or just in a neighbourhood of the unknown break-point [23]. Sometimes extra parameters modelling the smoothness of the approximating function are to be estimated [21] or selected *a priori* [11, 23]. Furthermore, the Gaussian model, which has been the only one investigated in these works, is necessarily requested in references [11, 23]. However, the change-point is no longer a parameter to be estimated.
4. In order to allow the regression function also to have continuous first derivatives, Pastor and Guallar [24] suggest assuming a linear effect before the break-point and a quadratic one after such a value. Such a linear-quadratic segmented model, however, implies relative risk increasing rather than constant after the break-point and a non-standard maximization algorithm has to be used to estimate the model. Simulations have shown the limiting distribution of parameter estimates is not well approximated by the Gaussian model, and high correlation among the estimates is inducted by the model.

5. Molinari *et al.* [3] use regression splines to estimate break-points (knots in spline terminology); a limit of this procedure is that the beta parameters are referred to the pseudo-variables inducted by the basis splines and so they do not measure the effect on the response [12]. Additional procedures, such as bootstrap techniques in reference [3], have to be used to get confidence intervals for break-points.
6. Bayesian MCMC methods have also been used recently. In a Bayesian approach no differentiability assumption is needed, but on the other hand, the computational effort might become rather heavy even with a simple model [14].

Thus, in short, most of the methods which have been reviewed seem to be inappropriate for at least one of the following reasons: (i) only single break-point detections have been investigated, and just a few have been employed in the case of multiple parameters; (ii) the break-points are not always treated as 'true parameters'; (iii) the model is constrained to a particular probability distribution of the response variable (often Gaussian or even logistic).

Furthermore, a common aspect of the methods mentioned above is the computational effort that possibly increases with the complexity of the model. The non-linear methods need starting values for all the parameters and the Bayesian approach by means of MCMC methods relies on a very large number of iterations. Therefore, although in principle some of them could be generalized, such extensions could not be immediate [19]. These non-standard procedures might be a practical limitation of modelling break-points in regression models.

4. THE MODEL

4.1. Fitting non-linear parameters in general

Let $h(z;\psi)$ be the non-linear term for the Z variable with parameter ψ in the generic regression model

$$g(E[Y]) = \eta(X) + \beta \times h(z;\psi) \tag{1}$$

Hereafter the predictor $\eta(X)$ may include any variable with a linear parameter, including some smoothing terms, say. $g(\cdot)$ expresses the link function for $E[Y]$, but in general it may represent the response in any regression model with linear predictor; because no difference exists among linear models in the following, both the response and the predictor will be omitted.

In general a multi-dimensional parameter is allowed, but just now it is assumed $\psi \in \Re$ for simplicity.

Following the idea from Box and Tidwell [25], also briefly mentioned in McCullagh and Nelder (reference [26], p. 379), it is possible to approximate the generic non-linear term by a first-order Taylor expansion around an initial known value $\psi^{(0)}$

$$h(z;\psi) \approx h(z;\psi^{(0)}) + (\psi - \psi^{(0)})h'(z;\psi^{(0)}) \tag{2}$$

where $h'(\cdot;\psi^{(0)})$ is the first derivative of $h(\cdot)$ in $\psi^{(0)}$.

If the parameter for $h(z;\psi)$ is β, it follows that the right side in (1) can be approximated by a fully linear new predictor, that is, $\eta(X)$ with additional terms

$$\beta \times h(z;\psi^{(0)}) + \gamma \times h'(z;\psi^{(0)}) \tag{3}$$

where $h(z;\psi^{(0)})$ and $h'(z;\psi^{(0)})$ are two new variates and $\gamma = \beta \times (\psi - \psi^{(0)})$, all depending on $\psi^{(0)}$.

Fitting the model with the new right side $\eta(X)+(3)$ yields ML estimates at each cycle for all parameters including $\hat{\beta}$ and $\hat{\gamma}$. Thus it follows that

$$\hat{\psi} = \frac{\hat{\gamma}}{\hat{\beta}} + \psi^{(0)} \tag{4}$$

is the updated estimate for the non-linear parameter ψ. The two new variates $h(z;\cdot)$ and $h'(z;\cdot)$ are reassessed, the model is refitted and a new estimate of ψ is gained; the process is iterated until possible convergence, which is not, in general, guaranteed (see reference [25] for details). As a result, when the algorithm converges, ML estimates for all the parameters in the model, including ψ, are obtained.

Inferences on $\hat{\psi}$ may be drawn by means of bootstrap, likelihood-based or Wald-type methods. In particular, in order to be able to use the simplest Wald statistics, the standard error of $\hat{\psi}$ may be obtained using linear approximation for the ratio of two random variables (that is, the delta method):

$$\mathrm{SE}(\hat{\psi}) = \{[\mathrm{var}(\hat{\gamma}) + \mathrm{var}(\hat{\beta})(\hat{\gamma}/\hat{\beta})^2 + 2(\hat{\gamma}/\hat{\beta})\mathrm{cov}(\hat{\gamma},\hat{\beta})]/\hat{\beta}^2\}^{\frac{1}{2}} \tag{5}$$

where $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot\,,\cdot)$ mean the variance and covariance, respectively.

4.2. Fitting segmented regression

A possible parameterization to model segmented relationship between the response and the variable Z is to fit the terms

$$\alpha Z + \beta(Z-\psi)_+ \tag{6}$$

where ψ is the break-point and $(Z-\psi)_+ = (Z-\psi) \times I(Z>\psi)$ being $I(\mathrm{A})=1$ if A is true. According to the parameterization (6) α is the slope of left line segment (that is, for $Z \leqslant \psi$), and β is the 'difference-in-slopes' parameter and $(\alpha+\beta)$ is the slope of the right line segment; thus if a break-point exists, $|\beta|>0$. Note the log-likelihood is not differentiable at $Z=\psi$.

The key in fitting segmented regression by means of the linearization (2) is that relevant first-order Taylor's expansion around $\psi^{(0)}$ holds exactly, provided that $\psi^{(0)}$ is the break-point:

$$(Z-\psi)_+ = (Z-\psi^{(0)})_+ + (\psi-\psi^{(0)})(-1)I(Z>\psi^{(0)})$$

where $(-1)I(Z>\psi^{(0)})$ is the first derivative of $(Z-\psi)_+$ assessed in $\psi^{(0)}$.

Therefore, by virtue of that told previously, the algorithm at each step s is:

1. Fix $\psi^{(s)}$ and calculate

$$U^{(s)} = (Z-\psi^{(s)})_+ \quad \text{and} \quad V^{(s)} = -I(Z>\psi^{(s)})$$

2. Fit the model with additional variates $U^{(s)}$ and $V^{(s)}$, namely

$$\alpha Z + \beta U^{(s)} + \gamma V^{(s)} \tag{7}$$

3. Improve the break-point estimate by (4).
4. Repeat the process until convergence.

The basic point to be noted here is that the non-linear and non-differentiable problem, namely the model (6), has been reduced to iterative fitting of linear models (7) throughout the variates U and V; the coefficient of U, β, models the difference-in-slopes as in equation (6) and the coefficient of V, γ, may be thought of as a reparameterization of ψ. At each iteration s, coefficient γ measures the difference between the two fitted straight lines (before and after $\psi^{(s)}$) at $Z = \psi^{(s)}$. Since with intersecting straight lines the gap is null, when the algorithm converges $\hat{\gamma}$ it is expected to be approximately zero, or rather non-statistically different from zero. Improvements in break-point estimation depend on such estimates through equation (4), that is, $(\psi^{(s+1)} - \psi^{(s)}) = \hat{\gamma}/\hat{\beta}$. When the algorithm stops and $\hat{\gamma} \approx 0$, there is no improvement in the break-point estimate and therefore the sth approximation is assumed the ML estimate, that is, $\psi^{(s)} \equiv \hat{\psi}$. Note that, given the meaning of β (that has to be greater than zero), this ratio never goes to infinity.

The algorithm may be shown to be exact, therefore it always will converge in a deterministic model or in situations with low variance, provided that the break-point exists, see Discussion.

The confidence interval for the break-point may be easily calculated by means of its Wald statistic; in a ratio of two random variables, such as equation (4), it is well known that as n increases and the denominator becomes more significant, a Normal distribution may be assumed (reference [26], p. 251; see also simulation results). Thus a 95 per cent Wald-based confidence interval, say, has extremes $\hat{\psi} \pm 1.96 \times \mathrm{SE}(\hat{\psi})$ where $\mathrm{SE}(\hat{\psi})$ may be calculated by formula (5). Note that the standard method to deal with ratios of random variables is through Fieller's theorem; however, it should be stressed that it works better than the delta methods if both numerator and denominator are significant [27], but in this context, since it is expected that $\hat{\gamma} \approx 0$, just the denominator $\hat{\beta}$ is significant and formula (5) should be preferred. Furthermore, if $\hat{\gamma} = 0$, formula (5) becomes $\mathrm{SE}(\hat{\psi}) = \mathrm{SE}(\hat{\gamma})/\hat{\beta}$ and $\mathrm{cov}(\hat{\psi}, \cdot) = \mathrm{cov}(\hat{\gamma}, \cdot)/\hat{\beta}$. These quantities may be easily obtained by fitting the new variate $V \times \hat{\beta}$ instead of V, taking V and $\hat{\beta}$ from the last iteration when the algorithm has converged. In this way approximate covariances of $\hat{\psi}$ with other parameters of the model are readily available from the output of the final model.

Of course the ideas discussed above to estimate the break-point of single Z will naturally apply to more segmented relationships with respect to different variables, and furthermore these can be easily extended to handle multiple break-points concerning the same single variable; this case is illustrated in the following subsection.

4.2.1. Multiple break-point parameter. Multiple break-points $\psi = (\psi_1, \psi_2, \ldots, \psi_K)^{\mathrm{T}}$ with respect to the same variable Z can arise in at least two cases:

1. The segmented relation is different among the levels $w_1, w_2, \ldots, w_K$ of some categorical variable W, and so there exists one break-point ψ_k together with right and left slopes for each group $k = 1, 2, \ldots, K$ (see Figure 1, right plot, for $K = 2$).

2. The relationship between the response and the single Z experiences several changes with respect to K change points (see Figure 1, middle plot, for $K=2$).

The latter may be thought of as a particular case of the former that is discussed below in some detail.

Assuming K-levels, a useful parameterization is given by

$$\psi = \psi_1 W_1 + \psi_2 W_2 + \cdots + \psi_K W_K$$

where $W_k = 1$ for observations belonging to group k and zero otherwise; ψ_k is the threshold parameter in group k. The non-linear terms in question depend on the interaction between the two variables Z and W, namely the K product-terms $\{Z \times W_k\}$:

$$\sum \alpha_k \{Z \times W_k\} + \sum \beta_k (\{Z \times W_k\} - \psi_k)_+$$

Expanding them in a first-order Taylor approximation yields $2K$ new variables at each iteration s: $U_k = (\{Z \times W_k\} - \psi_k^{(s)})_+$ and $V_k = -I(\{Z \times W_k\} > \psi_k^{(s)})$ for $k = 1, 2, \ldots, K$. Then it follows that

$$\sum_k \alpha_k \{Z \times W_k\} + \sum_k \beta_k U_k + \sum_k \gamma_k V_k \tag{8}$$

are the linear terms modelling piecewise regression with change-points depending on the categorical variable W. Successive approximations for the threshold parameters are given by $\psi_k^{(s+1)} = \frac{\hat{\gamma}_k}{\hat{\beta}_k} + \psi_k^{(s)}$ for every k.

Multiple change-points with respect to the same segmented relation are non-linearly modelled by $\alpha Z + \sum \beta_k (Z - \psi_k)_+$. According to this parameterization, α is the 'first slope', that is, when $Z \leqslant \psi_1$ and β_k is the difference-in-slopes parameter before and after ψ_k, that is, the difference between the $(k+1)$th and the kth slope. Then $\alpha + \sum_k^{\bar{k}} \beta_k$ is the slope for $\psi_{\bar{k}} < Z \leqslant \psi_{\bar{k}+1}$.

Assuming a multi-dimensional break-point ψ and $W_k = 1$ for every k, the (8) becomes

$$\alpha Z + \sum \beta_k U_k + \sum \gamma_k V_k$$

to handle multi-changes in single segmented relations.

Although in principle it should be possible to fit any multi-dimensional parameter, a few break-points (from one to three at most) are probably sufficient to handle several practical situations because the meaning of 'change-points' becomes rather questionable when their number increases [2]; in this case smoothing terms should be used instead.

5. SIMULATION STUDY

In order to assess the sampling distribution of the proposed break-point estimator, simulations have been carried out under different scenarios: different sample sizes (n), different break-point locations (ψ), and different difference-in-slopes parameters (β). Recently, these factors have been shown to be important in break-point detection in Gaussian models [28]. Given n, ψ and β, log-linear Poisson models have been generated with a standard uniform explanatory variable, namely

$$Y \sim \text{Pois}(\exp\{\eta\}) \quad \eta = 3.5 - 1.5Z + \beta(Z - \psi)_+ \quad Z \sim \text{U}(0,1)$$

Table I. Simulation results for break-point estimator (1000 replicates): descriptive statistics of empirical sampling distributions and summary from corresponding Wald-based confidence intervals.

Model	ψ	n	Sampling distribution summary				95 per cent CI summary†		
			Mean	Median	SD*	MSE*	%CP	AW*	WSD*
$\beta = 1.8$	0.50	50	0.501	0.505	0.922	0.085	88.3	2.91	0.85
		100	0.498	0.503	0.660	0.044	88.5	2.02	0.36
		500	0.499	0.499	0.271	0.007	90.5	0.92	0.07
		1000	0.500	0.500	0.184	0.003	92.8	0.66	0.03
	0.75	50	0.705	0.737	1.532	2.174	80.3	5.52	10.8
		100	0.718	0.743	1.201	0.155	83.1	3.20	1.80
		500	0.741	0.746	0.449	0.021	87.4	1.35	0.25
		1000	0.746	0.747	0.276	0.008	90.0	0.92	0.11
$\beta = 2.5$	0.50	50	0.497	0.500	0.657	0.043	88.5	2.04	0.38
		100	0.497	0.501	0.438	0.019	89.0	1.40	0.15
		500	0.499	0.500	0.174	0.003	93.3	0.65	0.03
		1000	0.500	0.500	0.123	0.001	93.9	0.46	0.02
	0.75	50	0.711	0.740	1.309	0.186	83.2	4.03	6.89
		100	0.728	0.743	0.841	0.076	84.3	2.24	2.69
		500	0.745	0.747	0.300	0.009	87.8	0.95	0.12
		1000	0.748	0.749	0.186	0.004	90.4	0.64	0.05

* Values $\times$ 10.

† Wald 95 per cent confidence intervals summary: %CP (per cent coverage probability); AW (average width of CIs); WSD (standard deviation of CI width). See text for details.

The parameters used in simulations reflect real situations. Count data are rather frequent in epidemiological studies and furthermore Poisson distribution shows a particular feature relating to break-point estimation, not involved, for instance, in Gaussian data, see below. The slope values (that is, approximately the log-relative risk) sound coherent with the explanatory variable; see, for instance, the estimates in Section 6.2 for the exposure variable '`DUST`' distributed roughly in $[0, 1]$.

The method discussed here needs starting values just for the break-point. Because in practice these may be easily obtained, for instance by smoothed scatter plot, there is no reason to start the algorithm with values too far from the exact solution, and therefore initial guesses $\psi^{(0)}$ have been randomly extracted by a uniform variate: $\psi^{(0)} \sim \mathrm{U}(0.35, 0.65)$ if $\psi = 0.5$ or $\psi^{(0)} \sim \mathrm{U}(0.6, 0.8)$ if $\psi = 0.75$. The random choice might avoid bias due to choice of starting values.

The performance of the break-point estimator has been evaluated by means of 1000 replicates for each model, through mean, median, standard deviation (SD) and mean square error (MSE); these quantities may be very useful to assess bias and variance reduction with respect to increasing sample size. Furthermore, in order to investigate the appropriateness of Wald-based confidence intervals (CI), the coverage probability (%CP) has been computed to measure the percentage of samples with CI including the true parameter. Also, to study the effect of n, ψ and β on the interval estimate of break-points, the average width of CIs (AW) and their standard deviation (WSD) have been calculated. Results are shown in Table I.

As one would expect, in general results get better when n increases; the estimators become unbiased and the empirical coverage level approaches the nominal one (0.95).

The findings agree with those of Julious in hypothesis testing problems [28]. Sample size, break-point location and difference in slopes influence break-point estimation, however the way these 'factors' influence the estimator is different.

With regard to unbiasedness, the sample size effect is substantially negligible when the unknown location of ψ is at the middle of the range of the exposure variable; in this scenario the impact of β is also unimportant. On the other hand, if $\psi = 0.75$, an acceptable small bias is met just for $n \geqslant 500$. Anyway, it appears the sample size is well able to smooth the differences among different ψ and β values, and to allow the estimator to be asymptotically unbiased.

Rather different is the effect of such factors on the variance of the estimator since differences in SD (and in AW) due to different ψ and β are rather evident, even if $n = 1000$. Therefore the practical situation of having small β and ψ near to the edge of the explanatory variable is expected to lead to wider estimated CIs, even with moderate or large sample sizes. However, as n increases, ψ is estimated more accurately (both AW and WSD decrease).

ψ and β have been shown to be important in break-point estimation, since they are the parameters affecting the V-shaped relationship. This is usually but not always true, however. For instance, as Poisson distribution is bounded, low counts might cause flat relationships, making parameter estimation, including break-point detection, rather difficult, independently of n, ψ and β. To give an insight, assuming a model with $n = 1000$, $\psi = 0.50$, $\beta = 2.5$ and intercept

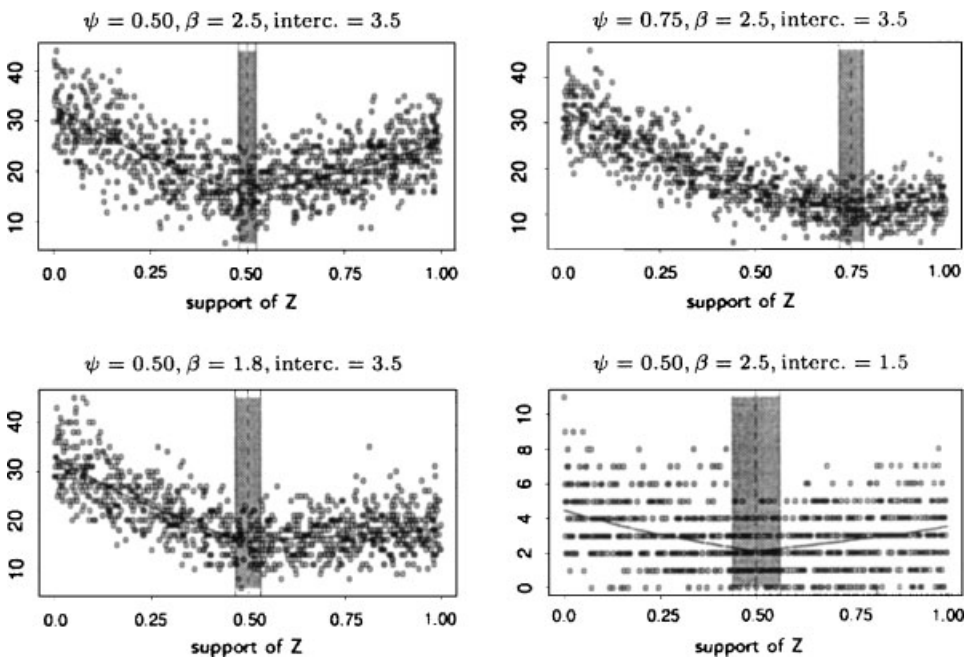

Figure 2. Four different hypothetical segmented relationships with $n = 1000$ Poisson counts. In each plot the average width (AW, based on 1000 replicates) of the 95 per cent CIs for the break-point is shown in grey colour.

equal to 1.5, the break-point estimator turns out to be unbiased, but with SD($\times 10$) = 0.373, approximately three times wider than the SD in the same model having intercept equal to 3.5.

Figure 2 shows for the same sample size and left slope ($n = 1000$ and $\alpha = -1.5$) the average width of the CI estimate under different parameter values; the narrowest AW is obtained with the 'gold-standard situation' ($\psi = 0.50$, $\beta = 2.5$ and intercept 3.5).

In conclusion, as it is obvious to suppose, the more clear-cut the segmented relationship, the easier the break-point estimation.

For the other parameters in the model, the relevant estimators were always unbiased with %CP very near to 95.0; this is not surprising at all, since at each step simple linear model (7) is fitted, and thus classical asymptotics hold perfectly.

Performance of the break-point estimator was also studied with binary data. For instance, using the parameters of reference [24] ($n = 1000$, whose 500 cases) the results were substantially unchanged; the estimator was unbiased (in mean and median) with SD($\times 10$) = 0.21 and coverage level 92.1 per cent. Therefore the large sample properties of the estimator seem to work even for binary data. Performances with Gaussian or continuous data are expected to work better, in particular when the variability is low or moderate.

6. APPLICATIONS

In this section the two examples presented in Section 2 are analysed and main results concerning break-point estimation discussed. Of course the presentations are necessarily incomplete and further analyses should be carried out, but here the goal is to stress that the method is suitable for different broken-line model problems.

6.1. Single break-point: Stanford heart transplant data

Figure 3 shows the penalized splines estimate for the hazard function with respect to age in a Cox regression model. Clearly it appears that the log hazard ratio is not constant with

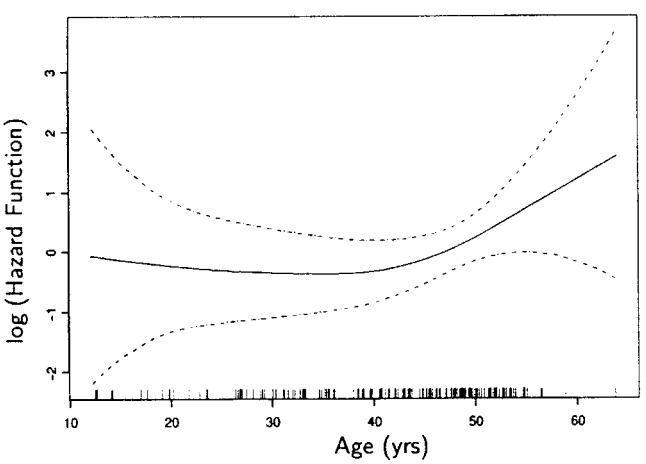

Figure 3. Non-parametric smoothing estimate (d.f. = 3) for the hazard function with respect to age.

Table II. Estimates from two Cox models for the Stanford heart transplant data.

Parameter	Model 1		Model 2	
	Estimate	SE	Estimate	SE
α	0.001	0.017	—	—
β	0.133	0.039	0.123	0.034
ψ	47.0	2.39	45.8	2.18

respect to age – subjects aged roughly more than 45–50 years have higher risk than the others. Such a non-linear relationship suggests fitting the piecewise term in order to estimate both the regression coefficients and the break-point. For comparison two models have been fitted: model 1 with three parameters to be estimated, that is, $\alpha\texttt{AGE} + \beta(\texttt{AGE} - \psi)_+$, and model 2 assuming $\alpha = 0$. Results are presented in Table II.

The change-point estimate in model 1 is equal to the one reported by Molinari *et al.* (47.0 [CI(95 per cent): 39.0–49.7]). This is not surprising since the two methods are believed to yield maximum likelihood estimates but with quite different procedures. Furthermore the method described in this paper allows us to obtain the hazard ratio for age less or more than 47 years, as well as the correlation between estimates; moreover, note that the model-based 95 per cent confidence interval for the break-point in Table II is slightly narrower with respect to the bootstrap-based one in reference [3] (9.37 versus 10.70).

Enforcing the left slope to be zero in model 2 allows us to reduce slightly the coefficients of variation (that is, SE over estimate), but does not modify the findings; subjects aged more than about 50 years have a higher significant risk. This confirms a J-like relationship between age at transplant and log hazard ratio.

Even if different starting values yield slightly different point estimates, they are associated with lower likelihood, and in practice the confidence intervals are always overlapped.

6.2. *Multiple break-points: Chronic bronchitis and dust concentration data*

Logistic regression models have been fitted in order to investigate associations between chronic bronchitis and explanatory variables `EXPO`, `SMOKE` and `DUST`.

Previous analyses used a main effect model assuming no interaction among the regressors, however a logit model with smoothing splines for `EXPO` and `DUST` has been fitted for both smokers and non-smokers revealing quite different patterns between the two groups, see Figure 4.

In the non-smoker group the `EXPO` and the `DUST` effects are virtually linear, with the `DUST` one substantially null. On the other hand, in the smoker group both the variables `DUST` and `EXPO` seem to have break-points. Likelihood ratio tests (not shown) have formally confirmed these impressions, that is, the effects of `EXPO` and `DUST` are linear in the non-smoker group and non-linear in the smoker group. Moreover, the plots suggest that the effect of `DUST` has a possible threshold value around 0.4, while the `EXPO` risk rises up to approximately 20 years, then remains constant and eventually increases after 35 years.

For this reason interactions $\texttt{SMOKE} \times \texttt{DUST}$ and $\texttt{SMOKE} \times \texttt{EXPO}$ have been included in the model. To get more interpretable parameters, the intercept has been omitted and both the

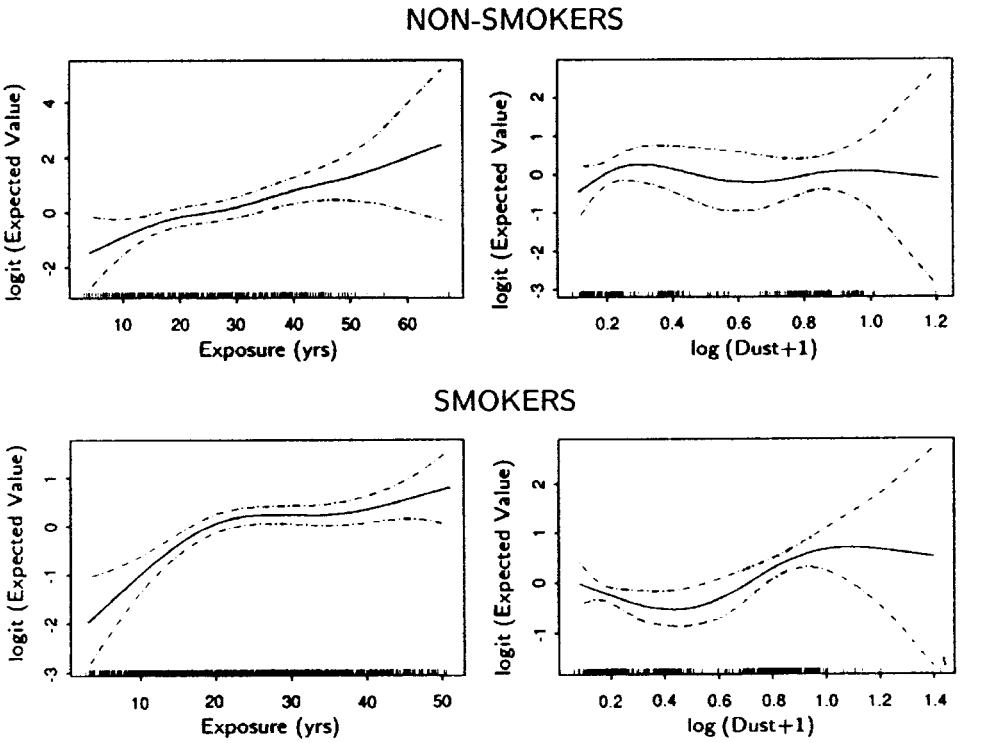

Figure 4. Smoothing splines estimates (d.f. = 3) for the effects of EXPO (left side) and DUST (right side) on the bronchitis in non-smokers (upper) and smokers (bottom).

two corner-point parameterized dummy variables of SMOKE have been fitted: SMOKE1 = 1 for non smokers and zero otherwise and SMOKE2 = 1 for smokers and 0 otherwise. Thus, for instance, the interaction SMOKE × EXPO is modelled through SMOKE1:EXPO and SMOKE2:EXPO. Three break-points have been estimated: one for DUST and two for the EXPO variable, both in the smoker group.

Therefore the overall ('iterative') linear predictor of the final model will be

$$\delta_1 \texttt{SMOKE1} + \delta_2 \texttt{SMOKE2} + \alpha_1^{\mathrm{E}} \texttt{EXPO:SMOKE1} + \alpha_2^{\mathrm{E}} \texttt{EXPO:SMOKE2} + \beta_1^{\mathrm{E}} (\texttt{EXPO:SMOKE2} - \psi_1^{\mathrm{E}(0)})_+$$
$$+ \beta_2^{\mathrm{E}} (\texttt{EXPO:SMOKE2} - \psi_2^{\mathrm{E}(0)})_+ + \alpha^{\mathrm{D}} \texttt{DUST:SMOKE2} + \beta^{\mathrm{D}} (\texttt{DUST:SMOKE2} - \psi^{\mathrm{D}(0)})_+$$
$$- \gamma_1^{\mathrm{E}} I(\texttt{EXPO:SMOKE2} > \psi_1^{\mathrm{E}(0)}) - \gamma_2^{\mathrm{E}} I(\texttt{EXPO:SMOKE2} > \psi_2^{\mathrm{E}(0)}) - \gamma^{\mathrm{D}} I(\texttt{DUST:SMOKE2} > \psi^{\mathrm{D}(0)})$$

$(\psi_1^{\mathrm{E}(0)}, \psi_2^{\mathrm{E}(0)})$ and $\psi^{\mathrm{D}(0)}$ are the break-point values, respectively, for EXPO and DUST in the smoker group, known from initial guesses or previous iterations; results are shown in Table III, model 1.

The findings on segmented relationships are doubtful; although the algorithm converges in less than 10 iterations, some issues are needed. With regard to EXPO, the second difference-in-

Table III. Estimates from two logistic models for the dust concentration data.

Parameter	Model 1		Model 2	
	Estimate	SE	Estimate	SE
δ_1	−3.145	0.421	−3.145	0.421
δ_2	−3.606	0.804	−4.222	0.721
$\alpha_1^{\mathrm{E}}(\times 10)$	0.532	0.131	0.533	0.131
$\alpha_2^{\mathrm{E}}(\times 10)$	1.541	0.408	1.673	0.507
$\beta_1^{\mathrm{E}}(\times 10)$	−1.784	0.468	−1.581	0.518
$\beta_2^{\mathrm{E}}(\times 10)$	0.685	0.548	—	—
α^{D}	−2.093	3.129	—	—
β^{D}	4.211	3.180	2.114	1.199
ψ_1^{E}	20.4	1.914	18.2	2.106
ψ_2^{E}	33.1	5.871	—	—
ψ^{D}	0.39	0.125	0.53	0.186

slopes estimate, $\hat{\beta}_2^{\mathrm{E}}$, is not significant and furthermore the 95 per cent Wald-based confidence intervals of ψ_1^{E} and ψ_2^{E} are overlapped: $(15.9; 22.8)$ and $(22.2; 44.5)$, respectively. Thus, although no formal test supports it, there is some evidence that ψ_2^{E} is unnecessary. Also the effect of `DUST` seems rather dubious because $\hat{\beta}^{\mathrm{D}}$ is marginally significant, however its left slope, $\hat{\alpha}^{\mathrm{D}}$, is statistically zero, and so it might be wise to drop it before making inferences on β^{D}.

In model 2 three parameters have been removed: β_2^{E} and consequently ψ_2^{E} for `EXPO:SMOKE2`, and α^{D} for `DUST:SMOKE2`. Now the effect of `EXPO` is much more clear; dust exposure approximately under 18 years has much greater impact on health, with respect to the estimated risk over 18 years (odds ratio 5.328 versus 1.096). This would suggest that adaptation to a dusty environment is important, because people working for a long time have less risk than the others working for short time; extra evidence is needed to support this hypothesis but this is beyond the aim of this paper. Dust concentration exhibits a threshold value around 0.53 (that is, 2.39 mg/m^3) above which the risk is just marginally significant (Wald-based p-value$= 0.078$). However, what should be stressed is the length of confidence interval being quite notable as compared to the range of `DUST`. This is not surprising at all, and actually it is also plausible given the rather wiggly-shaped relationship (see Figure 4 bottom right).

In conclusion, the results emphasize that risks are very different between non-smokers and smokers. In the former group there is no significant effect of dust concentration and the years of exposure have a negative constant effect. In the latter one, greater risks for exposure and dust associated with proper break-point values have been found, although the shape of the 'dose–response curve' for `DUST` needs much more investigation.

Results from previous analyses carried out on the same data set are available, but unfortunately estimates are not comparable for several reasons. First of all, interaction terms have been included in the model since the effects of `DUST` and `EXPO` turned out quite different between smokers and non-smokers; a break-point for `EXPO` in the smoker-group also has been estimated; measurement error in covariates has not been taken in account, and finally the sample size is slightly different from the previously published papers.

7. DISCUSSION

In this paper a simple method to fit segmented relationships in regression models has been shown; the non-linear and non-differentiable problem has been bypassed by simple iterative fitting of linear models.

The main advantage is flexibility in order to model more than one segmented relation and multiple break-points for the same variable, allowing dependence on categorical variables as well. The method is suitable for any regression model with a linear predictor, therefore classical GLMs as well as GAMs, survival models, ordinal response models (for example, proportional odds models) having segmented relationships with some explanatory variables may be fitted. Also, if requested, quasi-likelihood methods or robust approaches to account for possible outliers can be carried out as well. Furthermore, it is easily implemented and any current statistical software with its own language (for example, R, S-plus, SAS or Stata) may be employed using standard algorithms (Newton–Raphson, Fisher scoring) without any additional effort. This is a non-trivial advantage because, although other methods could be extended to fit more complex models, using *ad hoc* techniques to estimate non-linear models can turn out to be quite arduous, especially with several explanatory variables.

When the goal is just to get a point estimate of the break-point, some other methods discussed in Section 2 may be used to find the ML estimate, because the problem is just to find the maximum of the log-likelihood function, therefore the results should be the same (see the Stanford heart transplant example). However, unlike the method discussed here, they do not allow simultaneous inferences of all parameters of the broken-line model.

This paper, as many others, has not dealt with 'testing for the existence of threshold'. This problem has been studied by many authors [4, 8, 9], confirming that the distribution of the statistic test is quite complicated and also depends on the alternative hypotheses. However, in practice several tricks may be considered. First of all, smoothed scatter plots are a very useful means to detect segmented relations; if the plot does not show a possible change-point there is no reason to try to find it. With $\hat{\psi}$, a logical way to check the change-point is to test the differences between consecutive slopes [3]. For instance, according to the simple model (6), this may be accomplished easily by testing $\beta = 0$ using the Wald statistic $\hat{\beta}/\mathrm{SE}(\hat{\beta})$. Ulm [4] pointed out that the existence of a threshold (understood when the left slope is null) can be tested by one-sided hypothesis $\psi = z_{\min}$ versus $\psi > z_{\min}$. Of course, with break-points, where no constraint on the left slope is made, the maximum should also be considered (that is, the hypothesis $\psi = z_{\max}$ versus $\psi < z_{\max}$). Thus, as a rule of thumb, the confidence interval for the break-point should not include the extremes of Z, both maximum and minimum. Finally, comparisons among models by means of some likelihood-based criterion (for example, AIC) can be a useful guide [10].

The algorithm depends on the existence of a break-point. In general, if it converges, significant break-points are believed to exist, and eventually it is possible to test the difference-in-slope parameter as well as check the confidence interval. Otherwise, if the algorithm fails, it could be possible that a change-point exists but is not detected for a particular configuration of data (for instance, a small intercept in Poisson models); then one might not be able to say that the break-point does not exist, at least for small to moderate sample sizes.

A possible limit is that the success of the procedure can be dependent on the initial value $\psi^{(0)}$. This is a common problem for any non-linear model where, in general, starting values are needed for all parameters involved, and not only the break-points. This problem might be

emphasized by the fact that the profile log-likelihood for ψ is not always log-concave and so there may exist local maxima [28].

However, smoothed scatter plots can provide appropriate starting values for break-points and moreover several trials can be done to compare the sensitivity to different starting points; usually, if a break-point exists and the V-like relationship is clear-cut enough, the estimates from different starting values will be substantially the same; see the 'bronchitis–years of exposure' relationship.

Additional simulations (not shown here) carried out just for Gaussian data and moderate variance for one variable with a single break-point have demonstrated that the algorithm with a reasonable starting point always reached the exact solutions, with some negligible difference in the point estimate. Thus, some problem may arise just in a particular configuration of data; for instance, the relationship 'dust concentration–chronic bronchitis' appears quite unsteady (see the very large confidence bands in Figure 4 bottom right, in particular near to the possible break-point location). Therefore as a – probably correct – consequence, the confidence interval for ψ is very wide.

Some other difficulties can arise if there exists a high correlation among the estimates, in particular among the break-points. This can be easily verified by means of covariance matrix (for instance, in the dust bronchitis example the correlation between the two estimated break-points in model 2 is rather low, -0.13) but it is difficult to think about circumstances with many threshold parameters of interest.

Simulations have shown good performances of the proposed estimator, with empirical coverage level approaching the nominal one. Furthermore, as sample size increases, bias and standard deviation (and then mean square error) decrease, and the quantile-quantile plots look better (plots not shown). It might be possible deduce that the break-point estimator is consistent (that is, asymptotically unbiased with variance approaching zero as the sample size increases) and asymptotically normally distributed, although a formal proof appears quite desirable.

Therefore, in conclusion, the linear reparameterization as described in this paper seems a good approach to deal with broken-line models.

ACKNOWLEDGEMENTS

The author would like to thank to Dr Mariano Porcu and Professor Massimo Attanasio for reading the draft version of the manuscript and the editor and anonymous referees for their valuable comments which greatly improved the paper. The work was partially supported by grant MM13208412, 'Statistics in environmental risk assessment'.

REFERENCES

1. Stasinopoulos DM, Rigby RA. Detecting break points in generalised linear models. *Computational Statistics and Data Analysis* 1992; **13**:461–471.
2. Seber GAF, Wild CJ. *Nonlinear Regression*. Wiley: New York, 1989.
3. Molinari N, Daurès J, Durand J. Regression splines for threshold selection in survival data analysis. *Statistics in Medicine* 2001; **20**(5):237–247.
4. Ulm K. A statistical methods for assessing a threshold in epidemiological studies. *Statistics in Medicine* 1991; **10**:341–349.
5. Cox C. Threshold dose-response models in toxicology. *Biometrics* 1987; **43**:511–523.
6. Schwartz PF, Gennings C, Chinchilli VM. Threshold models for combination data from reproductive and developmental experiments. *Journal of the American Statistical Association* 1995; **90**:862–870.

7. Pires AM, Branco JA, Picado A, Mendonça E. Models for the estimation of a 'no effect concentration'. *Environmetrics* 2002; **13**:15–27.
8. Feder PI. The log likelihood ratio in segmented regression. *Annals of Statistics* 1975; **3**:84–97.
9. Beckman RJ, Cook RD. Testing for two-phase regressions. *Technometrics* 1979; **21**:65–69.
10. Ertel JE, Fowlkes EB. Some algorithms for linear spline and piecewise multiple linear regression. *Journal of the American Statistical Association* 1976; **71**:640–648.
11. Tishler A, Zang I. A maximum likelihood method for piecewise regression models with a continuous dependent variable. *Applied Statistics* 1981; **30**:116–124.
12. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: London, 1990.
13. Kunst AE, Looman CWN, Mackenbach JP. Outdoor air temperature and mortality in the Netherlands: a time series analysis. *American Journal of Epidemiology* 1993; **137**:331–341.
14. Gössl C, Küchenhoff H. Bayesian analysis of logistic regression with an unknown change point and covariate measurement error. *Statistics in Medicine* 2001; **20**:3109–3121.
15. Küchenhoff HK, Carrol RJ. Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statistics in Medicine* 1997; **16**:169–188.
16. Rigby RA, Stasinopoulos DM. Detecting break points in the hazard function in survival analysis. *Statistical Modelling* 1992; 303–311.
17. Küchenhoff H, Ulm K. Comparison of statistical methods for assessing threshold limiting values in occupational epidemiology. *Computational Statistics* 1997; **12**:249–264.
18. Vermont J, Bosson JL. François P, Rueff RC, Demongeot JA. Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine* 1991: **35**:141–150.
19. Küchenhoff H. An exact algorithm for estimating breakpoints in segmented generalized linear models. *Computational Statistics* 1997; **12**:235–247.
20. Hawkins DM. Point estimation of the parameters of piecewise regression models. *Applied Statistics* 1976; **25**:51–57.
21. Bacon DW, Watts DG. Estimating the transition between two intersecting straight lines. *Biometrika* 1971; **58**:525–534.
22. Griffiths DA, Miller AJ. Hyperbolic regression – a model based on two-phase piecewise linear regression with a smooth transition between regimens. *Communications in Statistics* 1973; **2**:561–569.
23. Tishler A, Zang I. A new maximum likelihood algorithm for piecewise regression. *Applied Statistics* 1981; **76**:980–987.
24. Pastor R, Guallar E. Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *American Journal of Epidemiology* 1998; **148**:631–642.
25. Box GEP, Tidwell PW. Transformation of the independent variables. *Technometrics* 1962; **4**:531–550.
26. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd edn. Chapman & Hall: London, 1989.
27. Neyman J. Contribution to discussion of the symposium on interval estimation. *Journal of the Royal Statistical Society*, *Series B* 1954; **16**:216–218.
28. Julious SA. Inference and estimation in a changepoint regression problem. *Statistician* 2001; **50**:51–61.

